Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **481,** 4030–4037 (2018)
Advance Access publication 2018 September 18

doi:10.1093/mnras/sty2536

Representative optical turbulence profiles for ESO Paranal by hierarchical clustering

O. J. D. Farley,[1]★ J. Osborn,[1] T. Morris,[1] M. Sarazin,[2] T. Butterley,[1] M. J. Townson,[1] P. Jia[1,3] and R. W. Wilson[1]

[1]*Centre for Advanced Instrumentation (CfAI), Durham University, Durham, DH1 3LE, UK*
[2]*European Southern Observatory, Karl-Schwarzshild-Str. 2, D-85748 Garching bei Muenchen, Germany*
[3]*College of Physics and Optoelectronics, Taiyuan University of Technology, Taiyuan 030024, China*

Accepted 2018 September 13. Received 2018 September 12; in original form 2018 July 18

ABSTRACT

Knowledge of the optical turbulence profile is important in adaptive optics (AO) systems, particularly tomographic AO systems such as those to be employed by the next generation of 40-m class extremely large telescopes. Site characterization and monitoring campaigns have produced large quantities of turbulence profiling data for sites around the world. However AO system design and performance characterization is dependent on Monte Carlo simulations that cannot make use of these large data sets due to long computation times. Here we address the question of how to reduce these large data sets into small sets of profiles that can feasibly be used in such Monte Carlo simulations, whilst minimizing the loss of information inherent in this effective compression of the data. We propose hierarchical clustering to partition the data set according to the structure of the turbulence profiles and extract a single profile from each cluster. This method is applied to the Stereo-SCIDAR (SCIntillation Detection And Ranging) data set from ESO Paranal containing over 10 000 measurements of the turbulence profile from 83 nights. We present two methods of extracting turbulence profiles from the clusters, resulting in two sets of 18 profiles providing subtly different descriptions of the variability across the entire data set. For generality we choose integrated parameters of the turbulence to measure the representativeness of our profiles and compare to others. Using these criteria we also show that such variability is difficult to capture with small sets of profiles associated with integrated turbulence parameters such as seeing.

Key words: atmospheric effects – instrumentation: adaptive optics – methods: statistical – site testing.

1 INTRODUCTION

In tomographic adaptive optics (AO), multiple wavefront sensors (WFSs) and deformable mirrors (DMs) are used to measure and correct the turbulence in the Earth's atmosphere over a wide field of view. This wide corrected field has made tomographic AO systems desirable for both current 8 m class telescopes (see e.g. Neichel et al. 2014; Esposito et al. 2016) and the next generation of 40-m class extremely large telescopes (ELTs; see e.g. Diolaiti et al. 2010; Hinz et al. 2010; Herriot et al. 2014).

In combining the off-axis WFS measurements to reconstruct the three-dimensional volume of turbulence projected from the telescope pupil through the atmosphere, some knowledge of the vertical distribution of the turbulence is required (Fusco et al. 2001; Vidal, Gendron & Rousset 2010). As such the performance of these systems depends on the optical turbulence profile, usually defined in terms of distribution of the refractive index structure constant $C_n^2(h)$ with altitude h. In particular, high-altitude turbulence where the spatial overlap between WFS measurements is small results in a degradation in AO performance.

The turbulence profile therefore plays a key role in the design of tomographic AO systems as they must be optimized for a particular observing site. As a consequence turbulence profiling forms a large part of site characterization studies (see e.g. Schöck et al. 2009; Vernin et al. 2011). These studies produce many measurements of the profile at a particular site. However, the majority of AO simulations used as part of the instrument design process (see e.g. Basden et al. 2007; Rigaut & van Dam 2013; Conan & Correia 2014; Reeves 2016) are Monte Carlo in nature and require long simulation times and many repeats of the simulation to produce results for a single set of atmospheric conditions. It is therefore not feasible to

★ E-mail: o.j.d.farley@durham.ac.uk

run simulations on many thousands of turbulence profiles to fully characterize AO performance for a particular site. Thus the large data set of measured turbulence profiles must be reduced to a small set that is in some way representative of the data set as a whole.

If the turbulence profile at a site were to show very little temporal variation, this task is relatively simple; the average integrated $C_n^2(h)$ values in each altitude bin for example would give a good approximation of the profile at all times. However for most observing sites the profile varies greatly on time-scales from minutes to seasons. In these cases such a method averages out features that are only present in a subset of the data, resulting in a profile that may never have been measured and is therefore not representative of the data set. An instrument optimized to such a profile would not perform as expected under real world conditions.

Here we put forward a method of obtaining a set of representative turbulence profiles at such a site by employing hierarchical clustering to provide a quantitative classification of profiles. This allows us to separate profiles with different structure and maintain the features in the profile whilst still reducing a large data set to a small set of profiles.

An example of a site with large variation in the structure of the turbulence profile is ESO Paranal, Chile. A 20-month long campaign using a Stereo-SCIDAR (SCIntillation Detection And Ranging) instrument (Shepherd et al. 2014) mounted on one of the auxiliary telescopes (ATs) has yielded a set of over 10 000 high-resolution (250 m altitude bins) measurements of the turbulence profile at Paranal (Osborn et al. 2018). We apply the clustering method to this data set to obtain a small set of turbulence profiles that we validate by comparing distributions of integrated atmospheric parameters. By ensuring the clustered profiles represent the data set in terms of these parameters we validate them in an atmospheric sense without reference to any particular AO system.

We can make the assumption that the free atmosphere turbulence at Paranal is similar to Cerro Armazones, the site of the planned European ELTs, since they are separated by only around 20 km distance and by around 500 m in altitude. As such this work is relevant to both sites.

In Section 2, we present an overview of hierarchical clustering and our method of extracting a small set of turbulence profiles from a large data set. In Section 3, we apply this method to the Stereo-SCIDAR data set from Paranal to obtain a small set of clustered profiles, with comparisons to other turbulence profiles for Paranal. Conclusions are in Section 4.

2 CLUSTERING

Cluster analysis allows underlying structure in large data sets to be ascertained by partitioning the data into subsets, known as clusters. There are many different ways to perform clustering on a data set but here we focus on hierarchical clustering (Everitt et al. 2011, chapter 4). We settle on this particular variety of clustering for two reasons. First, it allows easy switching and comparison of distance metrics, specifically non-Euclidean distance metrics that are particularly effective in this case. Secondly, the clustering can be visualized by the use of a dendrogram (see Fig. 1). At the lowest level we have each element in the data set represented by a vertical line, known as leaves. As we move up the dendrogram to larger distances elements are merged into clusters represented by the joining of two vertical lines into one. To define a certain number of clusters, we cut the dendrogram horizontally at a particular distance and count how many vertical lines (clusters) are intersected. In our case the dendrogram is most useful as a check that the clustering produces sensible results, especially when coupled with the data set ordered according to the leaves as also displayed in Fig. 1.

2.1 Distance metrics

The input to a hierarchical clustering algorithm is the distance matrix $\mathbf{D}$. For a data set of n observations of p variables (in this case $C_n^2\,\mathrm{d}h$ in p altitude bins), $\mathbf{D}$ is an $n \times n$ matrix whose components δ_{ij} represent the pairwise distances between all the observations using a given metric. The choice of the distance metric can have a large impact on the resulting clustering. The most commonly used metric is the Euclidean distance:

$$\delta_{ij}^{\mathrm{euc}} = \sqrt{\sum_{k=1}^{p} (\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})^2}, \tag{1}$$

where $\boldsymbol{x}_{ik}$ and $\boldsymbol{x}_{jk}$ represent the kth variables in two measurements of the turbulence profile $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ (Everitt et al. 2011, p. 49). This metric forms the basis of popular clustering algorithms such as K-means (Hartigan 1975). However for profiling data spanning several orders of magnitude in $C_n^2(h)$ the Euclidean distance proves to be very sensitive to outliers. As a result, clusters produced using the Euclidean distance tend contain a small number of extreme but very similar profiles, while assigning all other profiles (often over half the data set) to a single large cluster.

As an alternative, we found the cosine or angular distance to produce favourable results, defined as the normalized dot product:

$$\delta_{ij}^{\cos} = 1 - \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2}, \tag{2}$$

where $\|\boldsymbol{x}\|_2$ denotes the L2 norm of the vector $\boldsymbol{x}$. For positive data this metric is bound between 0 and 1. The cosine distance is less sensitive to outliers in our case and produces more reasonable clustering for turbulence profiles.

In calculating the distance matrix with profile measurement vectors $\boldsymbol{x}_i$ we have made the implicit assumption that all the components of the vector (altitude bins) are independent. This means that the height of the turbulent layer is not taken into account in the clusters and as such layers that are close in altitude are considered as similar in the distance matrix as layers far apart in altitude. This is not ideal especially since we are dealing with measurements with finite altitude resolution. We therefore modify the cosine metric as described in Sidorov et al. (2014). By introducing a $p \times p$ matrix $\mathbf{S}$ describing the similarity between vector components we obtain the soft cosine distance:

$$\delta_{ij}^{\mathrm{softcos}} = 1 - \frac{\sum_k^p \sum_{k'}^p S_{kk'} \boldsymbol{x}_{ik} \boldsymbol{x}_{jk'}}{\sqrt{\sum_k^p \sum_{k'}^p S_{kk'} \boldsymbol{x}_{ik} \boldsymbol{x}_{ik'}} \sqrt{\sum_k^p \sum_{k'}^p S_{kk'} \boldsymbol{x}_{jk} \boldsymbol{x}_{jk'}}}, \tag{3}$$

where both k and k' run through vector components. For $\mathbf{S} = 1$ this reduces to the cosine distance described in equation (2). The altitude resolution of the Stereo-SCIDAR is given by

$$\delta h = 0.5 \frac{\sqrt{\lambda |h - h_{\mathrm{conj}}|}}{\theta}, \tag{4}$$

where λ is the operating wavelength, taken here to be 500 nm, h_{conj} is the conjugate altitude of the imaging plane (for the Stereo-SCIDAR at Paranal $h_{\mathrm{conj}} = -3$ km), and θ is the separation of the double star used to compute the turbulence profile (Avila, Vernin & Masciadri 1997). We define each row k of $\mathbf{S}$ as a Gaussian with mean h_k and full width at half-maximum defined by equation (4). Each row is normalized such that all $S_{kk} = 1$. The widths of these Gaussians correspond very well to the response functions of the instrument

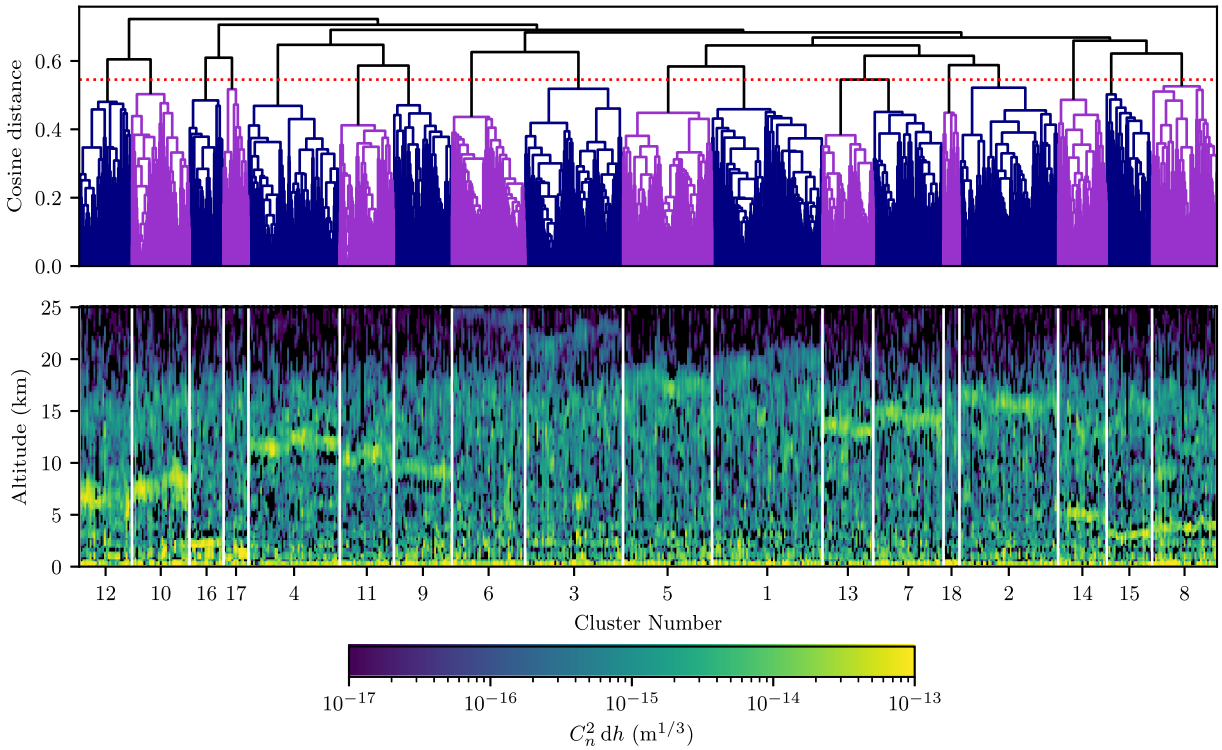

Figure 1. Upper: dendrogram representing average linkage agglomerative hierarchical clustering of the ESO Paranal Stereo-SCIDAR data set using the cosine distance metric. Branches below a cut-off distance of 0.55 (indicated by the dashed red line) are coloured alternately to indicate 18 clusters. Lower: the turbulence profiles in the data set, ordered according to the leaves of the dendrogram, with the partitioning into 18 clusters indicated by vertical white lines. Each cluster is assigned a number according to its size, with 1 being the largest cluster and 18 the smallest.

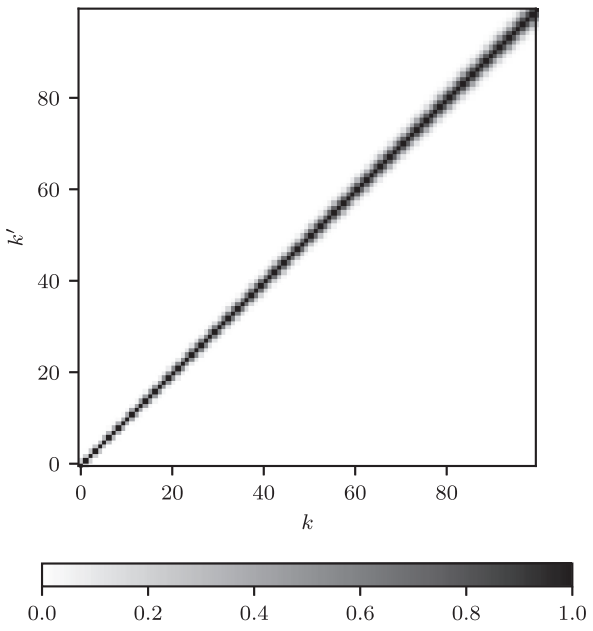

Figure 2. Similarity matrix **S** between altitude bins for the Stereo-SCIDAR at Paranal, using an average stellar separation of 12.5 arcsec, wavelength 500 nm, and conjugate altitude $h_{\mathrm{conj}} = -3$ km.

(Shepherd et al. 2014). The similarity matrix **S** used for the Stereo-SCIDAR data is shown in Fig. 2. This process ensures that the distance between profiles as defined by our metric takes into account the finite altitude resolution of the instrument.

2.2 Clustering process

The second choice that must be made in hierarchical clustering after the distance metric is the method of defining the intercluster distance or linkage. Here we use average linkage, where the intercluster distance is defined as the mean pairwise distance between the members of the two clusters.

A description of the process we employ to perform agglomerative hierarchical clustering is as follows.

(i) Compute pairwise distance matrix **D** for the chosen metric.
(ii) Merge the two closest elements.
(iii) Define the new distance from this cluster to the rest of the elements according to the chosen intercluster distance.
(iv) Repeat (ii) and (iii) until there are two remaining clusters that are merged into one representing the whole data set.

The clustering was performed in PYTHONusing the hierarchy module in SCIPY, which for average linkage clustering utilizes the nearest neighbours chain algorithm (see e.g. Müllner 2011).

2.3 Data pre-processing

The turbulence profiles contain many zero measurements. Usually these occur when turbulence in an altitude bin is below the sensitivity of the instrument but also can be a result of noise in the data post-processing pipeline. While it is tempting to treat all zero values as missing data and remove them from the analysis, this can have a profound effect on the calculation of distance between profiles. Thus we choose not to remove these zero measurements before clustering.

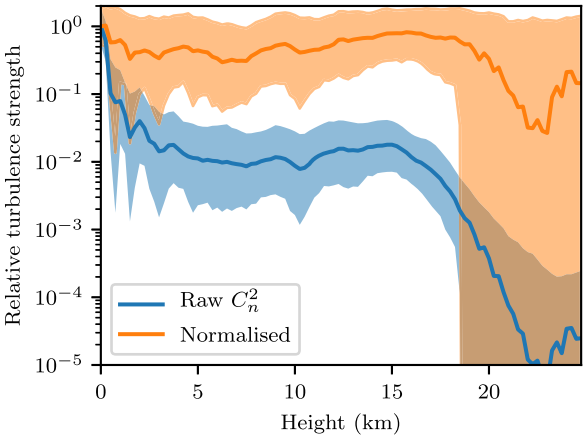

Figure 3. The effect on the median (solid line) and interquartile range (shaded areas) of normalization by dividing each altitude bin by its mean value. Turbulence strength is defined relative to the median value of the first (0 m) bin.

The dynamic range of C_n^2 measurements in the data poses a problem in clustering. The distance between profiles tends to be dominated by strong turbulence since these measurements can be up to 100 times stronger than weak or moderate turbulence (see Fig. 3). We are more interested in the significance of turbulence, i.e. whether turbulence is high or low relative to the average level of turbulence at a particular height. The C_n^2 measurements in each altitude bin are lognormally distributed but the censored nature of the data, where measurements below a sensitivity limit are recorded as zeros, means that we cannot log transform the data and perform the common procedure of subtracting the mean and dividing by the standard deviation for each altitude bin. Instead we find that simply dividing by the mean of each altitude bin is effective in 'flattening' the profiles, reducing the importance of strong ground layer bins and effectively increasing the importance of weak high layer turbulence such that turbulence at all heights is considered approximately equally in the clustering. The effect of this normalization on the distance matrix can be seen in Fig. 4. Note that the profiles are additionally L2 normed when the cosine distance is used.

2.4 Determining the number of clusters

We seek to cluster turbulence profiles until they are separated according to their structure, such that we can extract a profile from each producing a representative set of profiles. To quantify this we employ two metrics, the within cluster variance and the silhouette score.

We define the within cluster variance as the sum of the distances of the members of each cluster to the profile we extract as the centre of that cluster. We determine the distance with the same soft cosine metric used in the clustering:

$$W_N = \sum_{m=1}^{N} \sum_{i=1}^{n_m} \delta^{\text{softcos}}(X_{im}, X_m^*), \tag{5}$$

where n_m is the number of profiles in cluster m, N is the total number of clusters, the X_{im} are all the profiles in cluster m, and X_m^* is the centre of cluster m. The quantity W_N is analogous to the within cluster sum of squares that is minimized in K-means clustering, with the squared Euclidean distance substituted for the cosine distance and the cluster centroid $\bar{X}_m$ substituted for our more general cluster centre X_m^*. As we increase the number of clusters N, W_N will decrease rapidly at first with the gradient falling off as the clustering becomes less effective. It is at this point that we define the number of clusters, a technique known as the Elbow method.

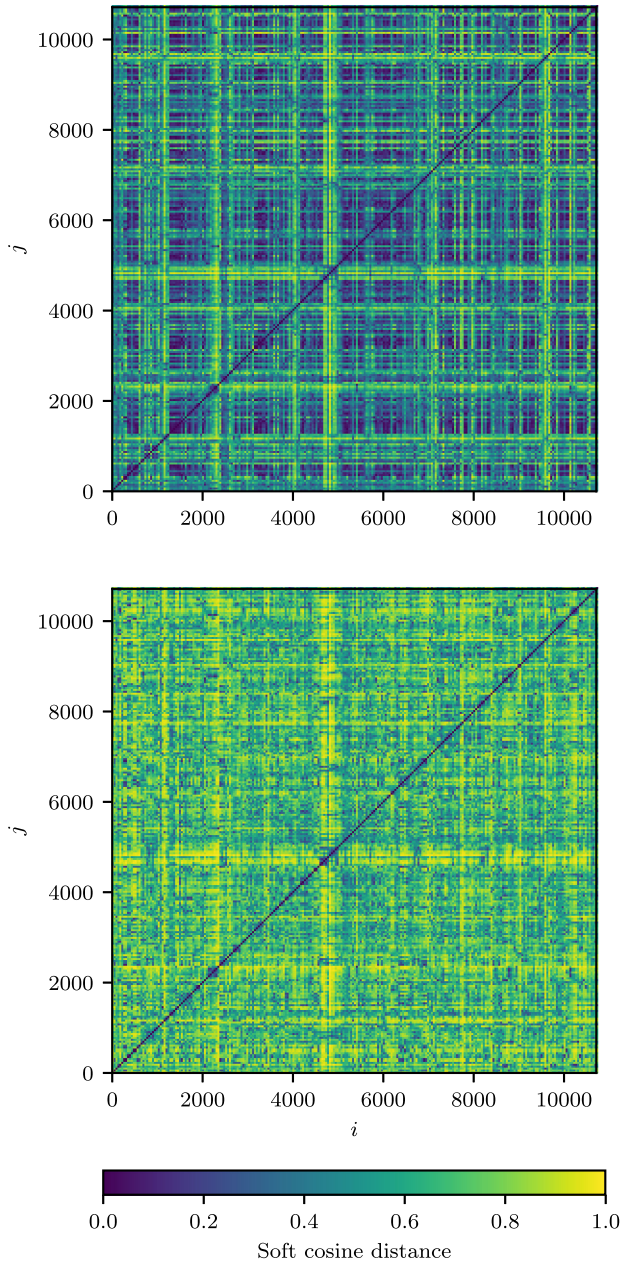

Figure 4. Pairwise distance matrices calculated using the cosine metric defined in equation (2) for the Paranal Stereo-SCIDAR data set of over 10 000 turbulence profiles. Top: raw C_n^2 measurements. Bottom: profiles normalized by dividing by the mean value in each altitude bin.

The second metric is the silhouette score (Kaufman & Rousseeuw 2005, chapter 5). This metric is defined for a single measurement i as

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \tag{6}$$

where a_i and b_i are quantities dependent on the distance matrix $\mathbf{D}$. a_i represents the average distance between measurement i and all the other members of the cluster i is assigned to. Conversely, b_i represents the average distance between i and all the members of

every other cluster. If $a_i > b_i$ resulting in $s_i < 0$, then this profile is on average closer to members of other clusters and is probably assigned to the wrong cluster. If $b_i > a_i$ then $s_i > 0$ and the profile is probably assigned to the correct cluster. A more positive silhouette score is therefore indicative of better clustering. s_i is by definition bounded in the range $-1 < s_i < 1$. By taking the mean silhouette score over all members of the data set $s = \frac{1}{n}\sum_i s_i$, we gain insight into the quality of clustering over all clusters.

These two metrics are chosen since, while not completely independent of one another, they incorporate distinct parts of the clustering process. The silhouette score depends solely on pairwise distances between profile measurements described in the distance matrix, whereas the within cluster variance also includes our chosen centre for each cluster $X*$. This allows us to draw a more robust conclusion as to the number of clusters in the data set.

2.5 Cluster centres

After performing the clustering and partitioning our data set we must extract a single turbulence profile from each cluster. The resulting profiles can vary greatly depending on the method used, so we present two methods and hence two sets of turbulence profiles here.

The simplest way to extract a profile from a cluster is to take an average of each altitude bin in a cluster. More specifically, we take the mean profile in our normed space, then unnormalize this profile and adjust it such that the integrated strength of the profile coincides with the median seeing for the cluster. This results in any features of the clustering common to all profiles in a cluster being retained while features belonging only to a subset of profiles will be averaged out as described earlier. The profiles thus produced will be an unrealistic but conservative description of the variability in profile and will represent the profile in the majority of cases.

Alternatively, we have already defined a metric that describes how well a profile fits into a particular cluster – the silhouette score. The profile in each cluster with the maximum silhouette score is therefore the best-fitting profile for that cluster according to our distance metric. In this way we can select an individual turbulence profile as the cluster centre. We therefore select the N profiles from the data set that represent the centre of each of the N clusters. These profiles will not be 'typical' in the sense that they represent the majority of measurements, but will describe a greater amount of variability that would also be useful for AO simulation.

3 APPLICATION TO ESO PARANAL DATA SET

We use the 2018A Stereo-SCIDAR data release described in Osborn et al. (2018). The data set consists of 10 691 turbulence profile measurements taken over 83 nights between April 2016 and January 2018. The profiles have 100 equally spaced altitude bins between the ground and 25 km.

The metrics for selecting the number of clusters are shown in Fig. 5. There is a clear peak in the silhouette score at 17–19 clusters. After 19 clusters the silhouette score drops off indicating that further clustering does not improve the quality of the resulting clusters. The within cluster variance in the average centre case shows no clear elbow but a transition from steep to shallow gradient at 15–20 clusters. In the single profile centre case, however, there is a clearer flattening of the gradient at 18 clusters, corresponding to the centre of the peak in the silhouette score. We therefore choose 18 as our number of clusters.

The magnitude of the silhouette score is only around 0.17 at the peak which is indicative of structure in the data that has not been captured in the clustering. Indeed we can see from the full set of extracted profiles shown in Fig. 6 that members of some clusters, especially those containing large numbers of profiles, are fairly inhomogeneous in structure. However, the clustering has for the most part selected and separated profiles with turbulence in strong single layers. This strong single layer is common to almost all profiles in a cluster. The lowest turbulent layers (e.g. clusters 14, 16, and 18) tend to be thinner and stronger whereas high layers (e.g. clusters 2, 4, and 5) tend to be more spread out and weaker. This may be an instrumental effect due to the reduction in native altitude resolution of the Stereo-SCIDAR with increasing height as described by equation (4) and included in the clustering by our use of the soft cosine distance. In total, clusters with significant high-altitude ($h \geq 10$ km) layers contain around 55 per cent of all profiles. We also have separated one ground-layer-dominated cluster (18) representing only 1.4 per cent of profiles. This propensity towards high-altitude turbulence is expected from atmospheric parameter statistics for this data: a median isoplanatic angle of 1.75 arcsec and fraction of turbulence below 600 m of 0.4 (Osborn et al. 2018).

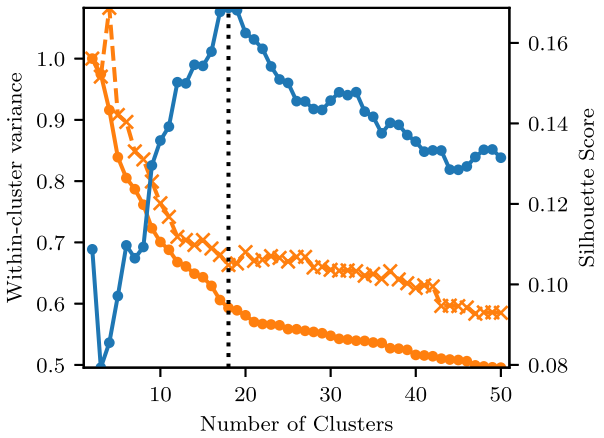

Figure 5. Within cluster variance (orange) and silhouette score (blue) for the Paranal Stereo-SCIDAR data set with increasing numbers of clusters. The two within cluster variance lines represent the two methods of defining the centre of a cluster: average (solid, circular markers) and single profile (dashed, cross markers). Within cluster variance in both cases is normalized to the value at 2 clusters. The dashed vertical line is at 18 clusters.

3.1 Comparison profiles

The most conventional way to reduce a large turbulence profile data base to a small set of representative profiles is to first bin the profiles according some integrated parameter, then take an average profile from each bin. The most common parameter used is the integrated strength (seeing), either measured from the profile itself or a contemporaneous measurement from a dedicated seeing monitor such as a Differential Image Motion Monitor (DIMM; Sarazin & Roddier 1990). This is the case for the ESO 35-layer profiles for Paranal (Sarazin et al. 2013), consisting of a profile associated with median seeing and four profiles associated with seeing quartiles. We also produce 18 profiles by binning the Stereo-SCIDAR data set into 18 seeing bins to provide a more equal comparison to our 18 clustered profiles.

In addition we compare to the good, high, and low profiles computed using the method defined in Sarazin et al. (2017). Rather than binning by the total integrated turbulence strength, the data set is split into three cases: good seeing, high-altitude dominated, and

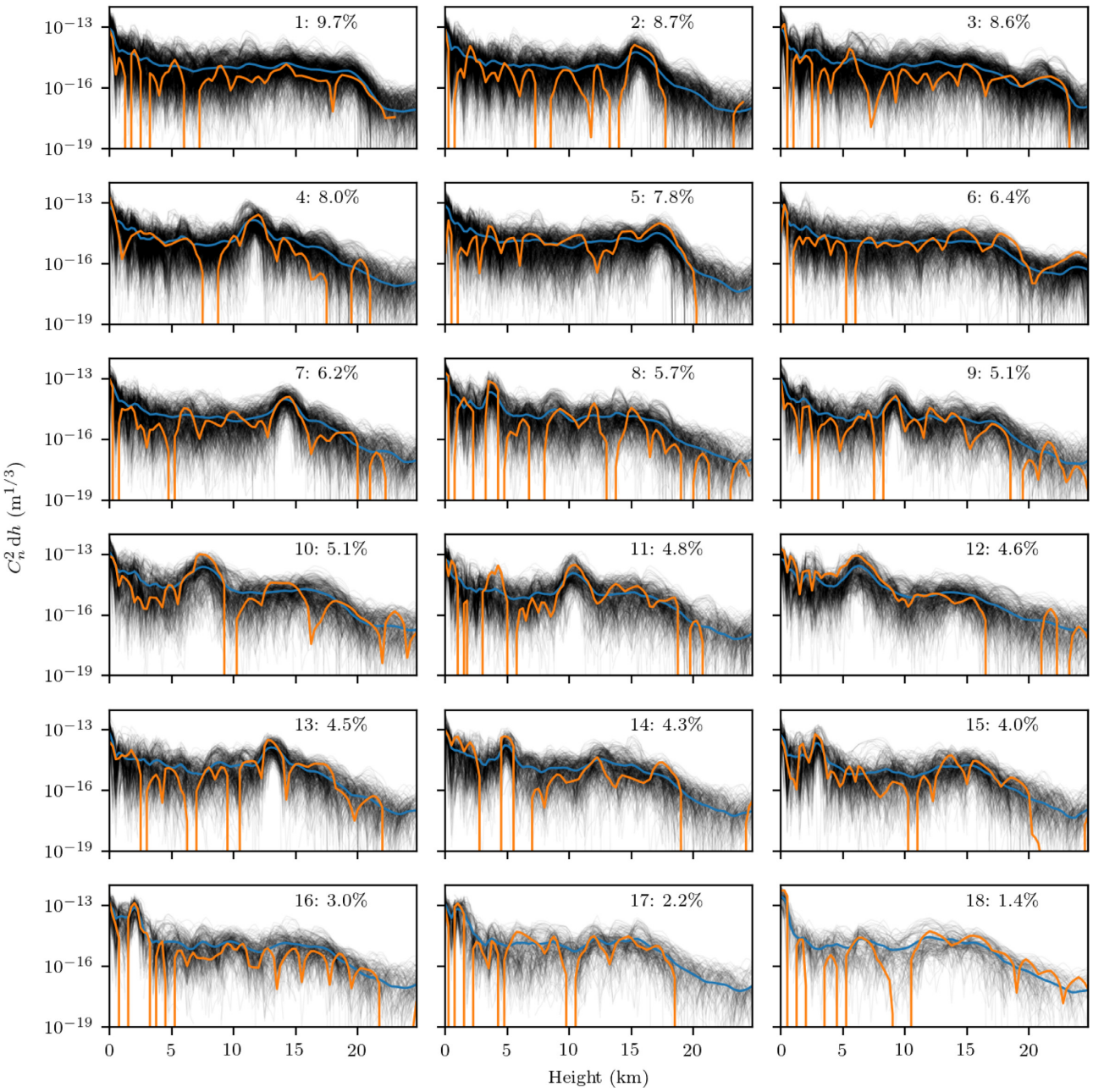

Figure 6. The set of 18 full atmosphere turbulence profiles for Paranal extracted through our hierarchical clustering method. Black lines represent every measurement of the turbulence profile in the given cluster. The two methods of obtaining the centre of each cluster are shown as blue (average profile) and orange (single profile) lines. Each cluster is numbered in descending order of the number of profiles in the cluster along with the percentage of all profiles contained in that cluster. Note that these profiles are not normalized.

low-altitude (ground layer) dominated profiles. The average from each of these cases is taken to produce three reference turbulence profiles for Paranal. We also include a profile 'all' defined as the average of all profiles in the data set.

3.2 Validation and comparison

Whether or not the clustered profiles represent the data set as a whole is a difficult question to answer since the concept of 'representativeness' can be defined in many different ways. The ultimate aim of this study is to produce a set of turbulence profiles that can be used in AO simulation with the knowledge that they reflect the variability in the turbulence profile seen in reality in some meaningful way.

The most direct method of validating the clustered profiles would be using fast analytical AO simulation (see e.g. Neichel, Fusco & Conan 2009) by comparing relevant AO metrics (e.g. tomographic error) over the data set to the clustered profiles. However, these metrics will depend strongly on the particular system simulated and are therefore beyond the scope of this paper.

In the interest of maintaining generality, rather than validating our profiles with AO simulation of one or several specific systems, we choose integrated atmospheric parameters as our metrics for validation and comparison to other profiles. While this general at-

mospheric validation will not necessarily agree with a tomographic AO simulation, these parameters serve as reasonable indicators for AO performance and are therefore a good compromise given the aforementioned sensitivity of AO metrics to the design of the particular system simulated. We choose the Fried parameter r_0 (Fried 1966) describing the strength of turbulence and isoplanatic angle θ_0 (Roddier 1981) describing angular correlation of turbulence, defined, respectively, as

$$r_0 = \left(0.423k^2 \int_0^\infty C_n^2(h)\,\mathrm{d}h\right)^{-3/5}, \tag{7}$$

$$\theta_0 = \left(2.91k^2 \int_0^\infty C_n^2(h)h^{5/3}\,\mathrm{d}h\right)^{-3/5}, \tag{8}$$

with $k = 2\pi/\lambda$ the wave vector of light considered (we take $\lambda = 500$ nm). We calculate these parameters for the entire data set and for our small sets of profiles and the results are shown in Fig. 7.

We can see that splitting the data set into 18 seeing bins and taking an average profile from each produces a set of profiles that by design fits very well with the distribution of r_0. However little of the variability in θ_0, a better indicator of the distribution of the turbulence, is described by these profiles. The ESO 35-layer median and quartile profiles behave in the same way. In particular, small values of θ_0 indicating significant high-altitude turbulence are poorly represented. The good, high, and low profiles provide a better description of the variability θ_0 but are slightly skewed towards larger values of r_0 indicating weaker turbulence. The 'all' profile lies in approximately the centre of both distributions as one would expect.

We include in the upper panel of Fig. 7 the distribution of integrated parameters for clustering with some different parameters to those presented above. We find that if we use the Euclidean distance instead of the soft cosine distance, the resulting clusters are heavily skewed towards smaller values of both r_0 and θ_0. Without normalization, the clustering produces profiles that better describe the distribution of θ_0 whilst being skewed towards larger values of r_0. Combining the soft cosine distance with the normalization described above (shown in the middle panel of Fig. 7), we produce profiles that accurately reflect the distributions of both parameters. However, the two methods of defining the centre of a cluster display very different results here. By taking an average profile for each cluster we produce a set of profiles whose integrated parameters are grouped tightly around the centre of the distribution for the data set. In the case of the r_0 distribution this is somewhat by design since we are not sensitive to changes in integrated strength (r_0) in our clustering, therefore we produce clusters whose individual distributions of r_0 follow approximately the distribution of r_0 for the entire data set. When we set the integrated strength of each of these clustered profiles to the median seeing for that cluster the values will tend to group around the median for the entire data set. In the distribution of θ_0, however, we see a similar tight grouping, with less of the bias towards larger values.

In contrast, if we take a single profile with the maximum silhouette score as our cluster centre, we produce a set of profiles that are spread more widely around parameter space. These profiles therefore describe more extreme variability. Again in the case of r_0 this is somewhat by design – since the clustering is not sensitive to r_0 we have essentially randomly sampled the distribution with 18 points, resulting in a wider spread around the parameter space.

Thus we have produced two sets of profiles that are both representative in different ways. Our average profiles are 'typical' since

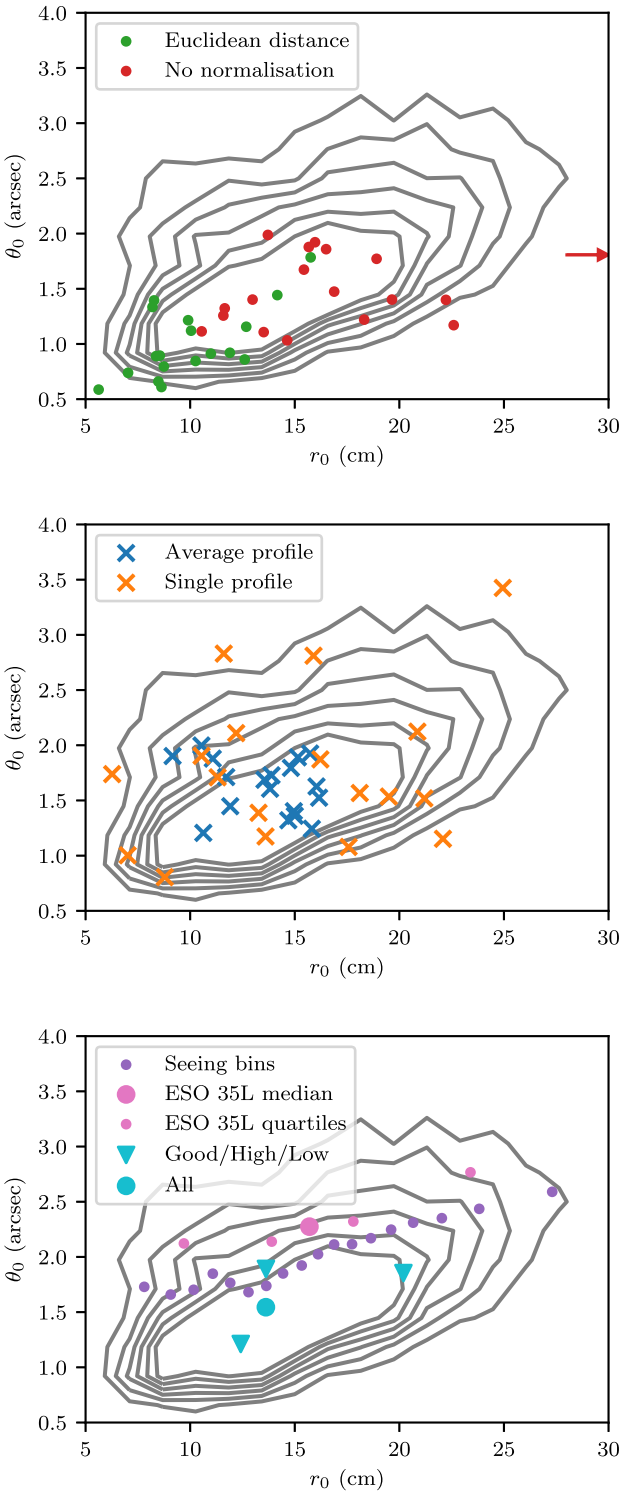

Figure 7. Distribution of integrated parameters r_0 and θ_0 for the entire data set (contours) and small sets of profiles. Upper: bad clusterings generated through our clustering method with suboptimal parameters. One outlier profile in the no normalization case with $r_0 = 33$ cm is indicated by an arrow. Middle: the two sets of 18 representative clustered profiles with cluster centres defined as average and single profiles. Lower: comparison profiles as discussed in Section 3.1.

they can be used to represent the profile most of the time. The single profiles are not typical since they represent a single measurement at a single time that is unlikely to represent the profile in the majority of times. However, these profiles exhibit more extreme variability in the atmosphere that would be useful in characterizing the performance of an AO system.

The turbulence profiles presented here are available on request to the author.

4 CONCLUSIONS

We have outlined a method for obtaining a small set of representative turbulence profiles from a large data set, where all steps of the process are informed by quantitative analysis of the clustering and resulting profiles.

We applied this method to the Stereo-SCIDAR data set from ESO Paranal, partitioning over 10 000 measurements into 18 clusters. We have used two methods to obtain the centre of each cluster resulting in two sets of 18 high-resolution full atmosphere turbulence profiles with 100 altitude bins between 0 m and 25 km. While the clustering has not preserved all the structural variation in the turbulence profile at Paranal, each cluster is dominated by a single strong turbulent layer, the height of which varies over the full range of altitudes.

Through analysis of integrated turbulence parameters it has been shown that the two sets of profiles are two distinct forms of 'representative' profile. Taking the average profile for each cluster results in typical profiles grouped around the centre of parameter space and representing the profile in the majority of cases. Conversely defining a single profile as the cluster centre produces a set of profiles that represent more extreme variability in the data set. Validation of these profiles for specific instruments using tomographic AO simulation remains for a future publication. Additionally, it would be possible to produce a set of profiles representative of the variability in profile for a particular instrument by performing the clustering on AO metrics relevant to that instrument (e.g. tomographic error).

Future work will focus on the temporal statistics of these clustered profiles, on both short time-scales of minutes to hours and longer seasonal time-scales. Analysis of seasonal variability in particular will require more data from the Stereo-SCIDAR to ensure statistically significant results.

More generally in the context of site characterization and monitoring, clustering methods can be applied not only to large data bases of turbulence profiles but also to any multivariate data (e.g. wind, humidity, and temperature) in order to extract small sets of representative conditions. Data from existing instruments such as AO telemetry or point spread functions could also be used either as input to the cluster analysis or as validation for representative atmospheric conditions.

ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Funding Council (UK) (ST/P000541/1). OJDF acknowledges the support of STFC (ST/N50404X/1).

Horizon 2020: this project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 730890. This material reflects only the authors' views and the commission is not liable for any use that may be made of the information contained therein.

PJ acknowledges the support of the National Natural Science Foundation of China(NSFC) (11503018, U1631133) and the China Scholarship Council (CSC).

This research made use of PYTHON including NUMPY and SCIPY (van der Walt, Colbert & Varoquaux 2011), MATPLOTLIB (Hunter 2007), and ASTROPY, a community-developed core PYTHON package for Astronomy (Robitaille et al. 2013). We also made use of the PYTHON AO utility library AOTOOLS (https://github.com/AOtools/aotools).

REFERENCES

Avila R., Vernin J., Masciadri E., 1997, Appl. Opt., 36, 7898
Basden A., Butterley T., Myers R., Wilson R., 2007, Appl. Opt., 46, 1089
Conan R., Correia C., 2014, Proc. SPIE, 9148, 91486C
Diolaiti E. et al., 2010, in Clénet Y., Conan J.-M., Fusco T., Rousset G., eds, Proc. Adaptative Optics for Extremely Large Telescopes (AO4ELT). EDP Sciences, Les Ulis, France, p. 02007
Esposito S. et al., 2016, Proc. SPIE, 9909, 99093U
Everitt B. S., Landau S., Leese M., Stahl D., 2011, Cluster Analysis, 5th edn. Wiley, New York
Fried D. L., 1966, J. Opt. Soc. Am., 56, 1380
Fusco T., Conan J.-M., Rousset G., Mugnier L. M., Michau V., 2001, J. Opt. Soc. Am. A, 18, 2527
Hartigan J. A., 1975, Clustering Algorithms. Wiley, New York
Herriot G. et al., 2014, Proc. SPIE, 9148, 914810
Hinz P. M., Bouchez A., Johns M., Shectman S., Hart M., McLeod B., McGregor P., 2010, Proc. SPIE, 7736, 77360C
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Kaufman L., Rousseeuw P. J., 2005, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York
Müllner D., 2011, preprint (arXiv:1109.2378)
Neichel B., Fusco T., Conan J.-M., 2009, J. Opt. Soc. Am. A, 26, 219
Neichel B. et al., 2014, MNRAS, 440, 1002
Osborn J. et al., 2018, MNRAS, 478, 825
Reeves A., 2016, Proc. SPIE, 9909, 99097F
Rigaut F., van Dam M., 2013, in Esposito S., Fini L., eds, Proc. Third Adaptative Optics for Extremely Large Telescopes (AO4ELT) Conf. INAF – Osservatorio Astrofisico di Arcetri, Firenze, Italy, p. 18. Available at: http://ao4elt3.sciencesconf.org/
Robitaille T. P. et al., 2013, A&A, 558, A33
Roddier F., 1981, Progress Opt., 19, 281,
Sarazin M., Roddier F., 1990, A&A, 227, 294
Sarazin M., Louarn M. L., Ascenso J., Lombardi G., Navarrete J., 2013, in Esposito S., Fini L., eds, Proc. Third Adaptative Optics for Extremely Large Telescopes (AO4ELT) Conf. INAF – Osservatorio Astrofisico di Arcetri, Firenze, Italy. Available at: http://ao4elt3.sciencesconf.org/
Sarazin M. S., Osborn J., Navarrete J., Milli J., Le Louarn M., Dérie F. J., Wilson R. W., Chacon A., 2017, Proc. SPIE, 10425, 104250B–10
Schöck M. et al., 2009, PASP, 121, 384
Shepherd H. W., Osborn J., Wilson R. W., Butterley T., Avila R., Dhillon V. S., Morris T. J., 2014, MNRAS, 437, 3568
Sidorov G., Gelbukh A., Gómez-Adorno H., Pinto D., 2014, Comput. Sistemas, 18, 491
van der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Vernin J. et al., 2011, PASP, 123, 1334
Vidal F., Gendron E., Rousset G., 2010, J. Opt. Soc. Am. A, 27, A253

This paper has been typeset from a TEX/LATEX file prepared by the author.